CHIVE PROGRESS REPORT (PHASE II)

September 18, 1963

Central Intelligence Agency
Office of Computer Services

SECTION I

General

This report discusses some preliminary ideas and assumptions which have evolved during Phases I and II of the DD/I systems study. As such it evidences a developing picture of the overall system beyond that presented in the back-up paper to the Phase I outline report.

Section 1 discusses the general system capabilities as presently conceived.

Section 2 is the General Development Plan and some general observations and assumptions.

Section 3 presents the Performance Specifications and a number of working assumptions in regard to performance objectives.

In Section 4 the implication of all of the above capabilities, objectives and assumptions in the resulting system configuration are discussed.

While much of what is said here is subject to revision and change, it is being furnished now to members of the CHIVE Evaluation Group in the hope that it will contribute to a better understanding of the CHIVE proposal.

1. General Systems Capabilities

This section covers the major functions that the system should or might perform. The capabilities are based on information gathered during the Fact Finding phase of the project and the firm conclusions drawn therefrom. Also included are some capabilities not yet made firm goals by management, but which might be required. Therefore, this section covers the maximum capabilities that will be expected of the system with the understanding that any of them may be relaxed.

1.1 Document Handling

The general requirement is to be able to "handle" all documents in use by the analytic offices which this system is to serve. Some relatively minor exceptions must be made, such as bound books for general use, certain graphic material, etc. For all material, the system must provide for the document to be indexed, both the index and the document (and/or a micro image of it) to be stored, and for retrieval of the index, the document, or a copy of it upon demand of an analyst or other user. Users must be able to request documents in terms meaningful to themselves although this does not exclude use of intermediary personnel who are specialists in file querying.

1.2 Information Handling

The system must also be able to handle factual data files which represent the final or intermediate results of analysis. Examples of such files are biographic summary files (as opposed to dossiers which are collections of documents), equipment characteristics files, gazetteers, etc. The principal differences between such information

files and document files are that information files: (a) are usually not written in natural languages--rather they are formatted, conveying information to the reader by virtue of position of words on the page as well as content, (b) are not copies of documents--they are end products--they contain "answers" to questions, and (c) may be searched directly by computers whereas document files are usually searched only through the medium of an index file.

"Handling" information files implies providing facilities to users to permit creation of new files, erasure of old files, changing of files, storage of files, adding information to files, retrieving information on a selective basis, or retrieving an entire file.

1.3 Centralized System Control

In order to improve service to users, the system should allow them to query all information in the system from any query point. A query point is the physical location at which an analyst presents his information needs directly to the system or to a specialist in querying. This capability speeds the analyst's search for information, reduces the amount of time and effort he must expend, and raises his confidence in the system. The implication to the system is that all material to be handled, regardless of source or classification, be indexed in an internally consistent manner, that a centralized index be maintained or that decentralized information stores be available to all query points, and that any physical access point be able to provide the same information to an analyst. (For security reasons, some access points

may not be able to provide delivery but the analyst there should be able to direct searching of an all-source file.)

1.4 Specialized Files

Many individual analysts or analytic organizations prefer to be able to maintain their own files, in addition to the centralized files. The reasons are generally two: (a) faster access, (b) analysts are not always sure enough of the information to want to submit it to a central point. The system should have the capability to permit such individual preference without prejudicing overall performance. For example, there is no reason why an individual may not retain a document in his desk so long as a copy is also in the central system. This gives everyone access to it. Information files can be similarly treated. Analysts may retain personal information files, may also store them in the system where the computer can be used for statistical analysis of the files or other calculations, or the analytic office may choose to enter an "official" version of a file for general use but retain a "hunch" file in their own office. It is also possible to store an information file in the system but deny access to all but the proper analytic office.

1.5 Document Dissemination

The system should have the capability to determine the dissemination of incoming documents on the basis of the document index and of analysts' statements of interest. In establishing this requirement, it is understood that implementation of automatic dissemination

may be made on a piecemeal basis, that is, document class by document class and organization by organization.

1.6 System Response Times

The time required by the system to respond to a query, to disseminate new information, or to sotre new information in a file should be such as not to delay analytic offices in the performance of their tasks. That is, analyst schedules should not be held up by delays in posting, disseminating, or retrieving information. In the case of queries submitted by analysts involved in long term research projects, this requirement places little pressure on the system. In the case of analysts engaged in analysis of rapidly changing, day-to-day situations, this is a very significant requirement.

2. General Development Plan

The project plan as presently established calls for a four-phase development:

Phase I: System Requirements Study.

This effort (which led to the System Capabilities shown in Section I) studied and evaluated user requirements for an IR System, and developed a planned program for design and implementation of such a system. This study was completed in June, 1963.

Phase II: System Design.

This effort was begun in July, 1963 and calls for the design of the IR system.

Phase III: Initial System Implementation.

This effort is planned to start in July 1964 for a period of about one year and will result in the implementation of the initial segment of the full system.

Phase IV: Expansion to Full System.

This effort will be concerned with gradual expansion of the system implemented in Phase III to cover the full input load.

2.1 Initial System Objectives

The basic objective of the initial system is to establish a small scale mechanical structure of the eventual system in a limited, controlled environment. This system will be designed with the performance specifications of the full system in mind for application to a limited area. One of the basic premises of the Initial System, however, will be that expansion in terms of added sources and increased performance will not require substantial redesign of programs, methods of operation or equipment configuration. This will require that the initial system have as many features of the full system as possible, subject to economic justification.

2.2 Role of the Contractor

As recommended in [redacted] of 3 June 1963 and agreed STATINTL to in the contract dated July 1, 1963, the contractor is to assist the government in the System Design effort by the performance of certain stated tasks.

These tasks are grouped under two major areas, Systems Engineering and Program Design and are described in detail in the contract.

During the performance of these tasks the contractor has

developed the following two sections (3 and 4) as a partial result of our efforts to date. These sections are intended to present performance specifications and certain equipment implications which necessarily result from the adoption of these specifications as working hypotheses.

2.3 System Cost Considerations

It is well to note at this point, that as yet no maximum ultimate system cost has been established. However, no inference has been made that such a ceiling does not exist.

The design group has been using as a working assumption that development of such a ceiling would be primarily based on a "cost-for-total-capability" basis rather than on such other measures as a "cost-per-document indexed" or " cost-per-query-answered."

3. Performance Specifications

3.1 Document Input

3.1.1 Volume

Initial System -- the initial system as defined in Section 2 will be an implementation of the full system concept on a limited volume of document input. Although the actual method for limiting this volume is still under consideration by Agency management, a working assumption was made that the volume of documents to be considered for input in one year would be 60,000.

Full System -- the full system should control current documentary input. The estimate of yearly document input volume is 1,000,000.

Growth -- in order to perform the system engineering and program design tasks, a range of growth rates from initial to full system were assumed. (Refer to Figure 3-1.) It is felt that the earliest possible date of the initial system implementation is January, 1965 or six months following the completion of the Phase II system design. The earliest possible date for implementation of the full system is considered to be July, 1968. The latest dates that are reasonable to consider for implementation of the initial and full systems are assumed to be July, 1966 and July, 1974 respectively.

Again, for planning purposes, a linear growth was assumed from the document rate for the initial system to the document rate of the full system, and the earliest and latest growth curves were plotted according to the range of implementation dates stated above. The area

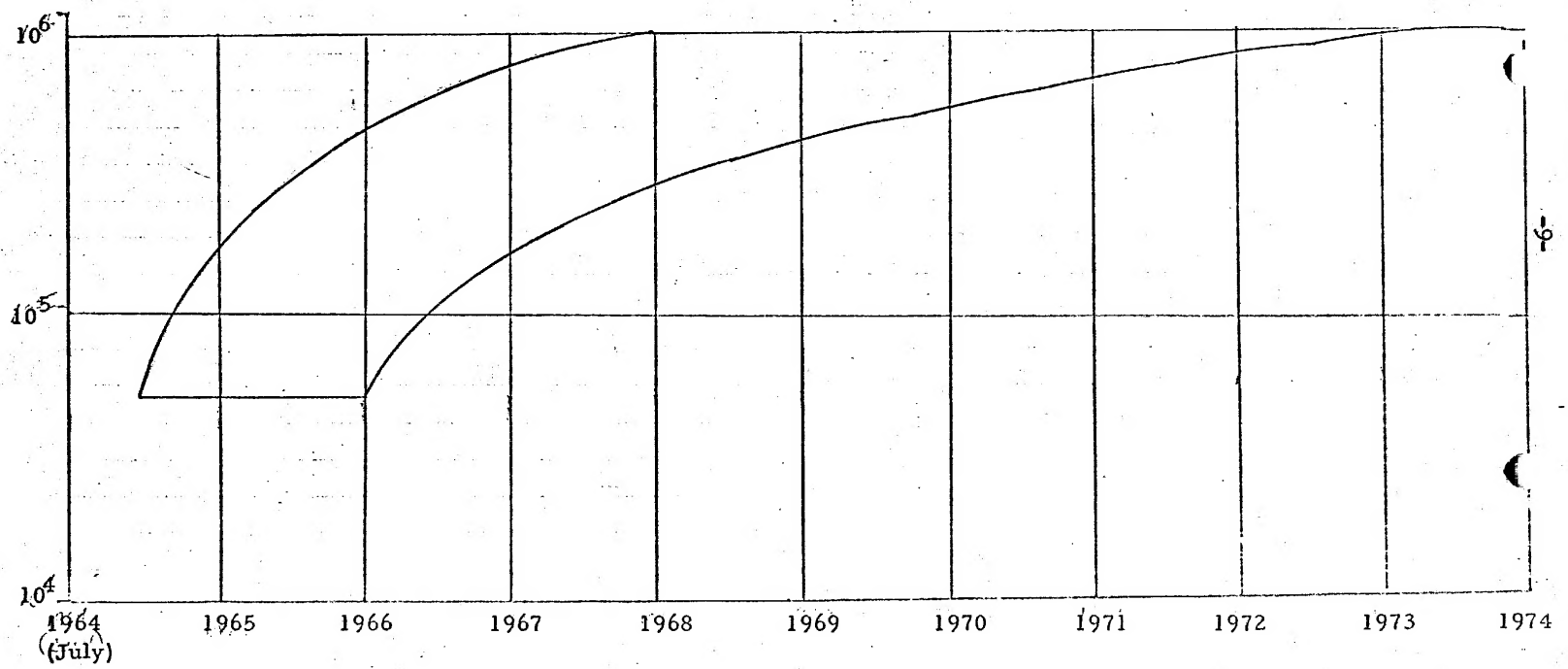


FIGURE 3 - Input Rate in Documents/year

bounded by the two curves represents the range of growth rates which are being considered.

3.1.2 Types

All types of documents that are now input to the user organization are being considered as input to the system. It is realized that there are variations in respect to such things as security classification, inclusion of pictures and graphics, quality of print, reliability of information content, machine readability, etc.

3.2 Document Indexing

3.2.1 Methods

The traditional methods of indexing may be roughly classified as follows:

All Symbolic -- all the index entries are completely coded fixed field terms, ISC codes and modifiers.

Controlled Keyword -- index entries contain keywords but only those appearing in a prepared thesaurus.

Free Keyword -- any keywords may be used in the index.

Free Keywords with Links -- same as the Free Keyword system except that linkage is permitted between keywords to indicate a phrase or relationship between words.

Phrases -- full phrases from the documents are used as the index so that to some degree the context in which a keyword was used may be determined on retrieval of the phrase.

Abstract -- a concordance of the abstract is used as the index giving additional ability to determine context.

Full Text -- a concordance of the entire text is the index to a document. Ability to determine context is maximized.

3.2.2 Indexing Rates, Personnel, Cost

The indexing rate will vary with the chosen indexing method. Assuming either the full keyword with linking or the phrase methods, the indexing rate is estimated at approximately 30 minutes per document per indexer. Of this time about 20 minutes will be devoted to reading and the remaining 10 will be devoted to indexing operations such as keyword selection and transcription.

The number of indexing personnel to support the proposed system grows from about 20 for the initial system to 300 for the full system. Again, as in document input growth, the build up of personnel may be rapid or slow depending on the implementation dates chosen as goals.

Manpower costs for indexing are assumed to be $7,000 per year per indexer. This is the only cost which will be considered as part of the indexing operation. Transcription costs will be covered as part of the input operation.

3.3 Information Files

2.2.1 Authority Files

This includes the dictionaries, name files, organization subordination files, etc. There is an assumed requirement for mechanized storage of these files. They should be changed or updated immediately when necessary and be available to indexers and analysts for queries. In addition, there is an anticipated requirement for periodic

printouts to be used at the remote indexer and analyst offices.

3.3.2 Fact Files

These files contain summary, factual information, each item of which is identified by labeling (tagging) or by its position in the files. They are not considered to be directly related to documents. The information for these files may be derived from analysis of documents which have been retrieved from the system -- but analysis is essential to forming fact files. These fact files are considered a necessary part of the system in order to respond to queries for information and in order to aid in the processing of queries for documents.

3.3.3 Size of Files

The exact format and number of these information files is not known at this time. It is assumed for planning purposes that the information files consume approximately 1/3 of the storage space consumed by the document index and inverted index (term) files.

3.4 File Changes

There exists a requirement to be able to change any of the system files. Such changes are necessitated by:

1) Errors -- These may be transcription errors, spelling errors, or indexer/analyst errors resulting in information entering the system which may be in error.

2) Changing environment -- The information files, in particular, are dynamic in nature. The attributes and the entries change rapidly. Such things as organizational changes, occupational changes, travel, etc. must be

reflected in the information files.

3) New information -- File entries are often made with a very limited amount of information. As more information is gathered it is often the case that documents have been wrongly classified or false entries have been made to information files.

3.5 Query Types

3.5.1 Information Query

Queries to the information files may be made by indexers or, ultimately, by analysts. The queries may be aimed at retrieving information to aid them in the performance of their duties or to aid in the formulation of information or document queries or index records. The latter functions are referred to as conversational querying. An indexer should have the capability of asking short information retrieval questions in order to:

1. Determine how a subject has been indexed before,
2. Identify a word or phrase, and
3. Determine what is already on file to avoid redundant entry.

The analyst may not know how to phrase his question to yield the desired result from the system. He should have the capability of asking a series of short questions, each question being predicated on the response to the last question. Thus, he may improve his original query avoiding retrieval of irrelevant material or non-retrieval of relevant material, or both.

3.5.2 Document Requests

Handling of queries resulting in retrieval of documents is a definite requirement of this system. Document queries may result in many pages of document output. In order to reduce the analyst reading problems and machine processing time some intermediate outputs may be desired:

1. The document index record may be delivered in response to the document query. The analyst may then use the index as an abstract of the document and accept or reject on this basis.
2. Microimages may be transmitted to the analyst for viewing. Thus, another screening may take place before ordering hard copy.

3.5.3 File Interaction

Queries should make use of all information available in the system regardless of the basic file being queried. For instance, a document query may be expressed as the intersection of a number of terms. If any of the terms was not in the document index, the document would not normally be retrieved. However, if additional information can be supplied by the information files for document query, the document may be found to answer the question.

3.5.4 Cyclic Queries

Queries should be capable of initiating a series of searches each yielding an intermediate output which is used as input to the next search. Thus, with a very limited amount of information, the

analyst may initiate a complex query.

3.6 Query Rates

3.6.1 Information Queries

Requests for retrieval from information files may occur at a rate as high as 3,000,000 per year. These queries may be from indexers or analysts, but so high a figure would only be reached by permitting and encouraging indexers to query while indexing.

3.6.2 Document Requests

Queries resulting in document retrieval should occur at an approximate rate of 100,000 per year. Each query may result in the retrieval of a number of documents. Thus, this figure is not the same as the number of individual image retrievals.

3.7 Response Time

3.7.1 Information Queries

Requests for information should be answered in one to two minutes. In using the conversational mode, it is important that once the first query has been entered the responses and additional queries proceed rapidly and without interrupt.

3.7.2 Document Queries

Response time required for document queries may be split into two categories, less than 15 minutes, and eight hours or more. The first category is assumed to be the requirement of about 25% of the document queries. The remaining 75% will have the less demanding requirement.

4.1 Input

4.1.1 Entry of Document Indexes

Input, in the context of this Section, refers to digital data to be stored in information or index files. Two special problems are created by the requirement to handle one million documents a year. First, the cost of transcribing indexes of these documents, by traditional means, is excessive. The very number of people--keypunch and verifier operators with their supervisors and administrative support--presents a large administrative problem. Second, error control, in so large an organization will be very difficult and conceivably could be such as to totally block successful use of the system. That is, unless an effective means for detecting and quickly correcting errors (e.g., misspelling a name in index record) can be found, file quality can quickly become degraded and the proportion of effort spent by both people and machines in reacting to errors can become excessive.

To attack these problems, we have looked at the basic nature of the indexing transcribing function. For an indexer to transcribe his data onto a data form, using careful block lettering, takes about as much time as typing the same information. If, however, the data were transcribed using a tape-generating or remote-input typewriter, the keypunch operation could be avoided with a sizeable savings in input preparation cost. As a bonus, the elapsed time to get new information entered into the system is dramatically reduced. An even more significant bonus, however, is that by directly connecting these transcribing devices to a computer an excellent error control technique is provided; that is,

computer detection of errors and the immediate feedback to the indexer. The advantages of quick feedback to the indexer are two:

(a) The indexer is presumably still working on the same document when the feedback is received and has the pertinent data fresh in his mind. Since re-reading of the complete document is not necessary, an extremely time-consuming task can be eliminated.

(b) The indexer, perhaps second only to the author, is the person best qualified to rectify an error in indexing for this will often require understanding of the subject area as well as the context of the document.

This procedure applies, of course, only to machine detectable errors. These errors are: spelling (e.g., use of a dictionary can detect when an unknown word is used), illegal classification codes, illegal combination of tags and/or keywords, format errors, and possibly unusual combination of subject classes and keywords which a computer might reject for further checking. Actual content review must be performed by humans, as now, either by full review of all indexes by supervisors or spot checking.

Where machine readable input is available, such as Teletype tape, measures can be taken to further speed the index transcription process. In this situation the indexer could be relieved of the necessity to transcribe words and phrases, instead having only to identify the word or phrase to be selected. Presuming a copy of the text will be made available to the computer, such identification

may be performed by entering the word number rather than the actual word into the computer. Net gain should be about half the time the indexer would normally spend in actual transcription.

4.1.2 Querying

Again, because of the large volume of input and number of documents and records kept in storage, a meansmust be provided to give users good quality responses to their queries without taking too much time. The method chosen is to provide for conversational querying -- a technique whereby the requester can scan files, determine what is in the files related to his query, and, if desired, change the question again. Reasons for changing the question might be: insufficient data on a given subject dictates use of a broader question; too much data requires a narrower question; or a completely different subject approach might be suggested by file records retrieved in response to the original query.

It is important to realize that conversational querying not only gives an analyst the positive advantage of improving his query rapidly, but also prevents the degradation of system performance that results when users ask overly broad questions to "play safe". There is no need for such action under this concept. If the user is dissatisfied with his first results he has only to try again--he is not subject to a waiting period of several days to do this.

The purposes, to repeat, of conversational querying are to permit rapid resolution of ambiguities,make associations between subjects and index terms, and eliminate redundant information. This

process can be accomplished to some extent either by the requester upon retrieval or by the indexer. That is, given the same capability to make rapid queries to the file, the indexer can resolve ambiguities and make associations between terms. This can result in more accurate, more compact files; hence, better retrieval. Whether the process is done at the retrieval end of the cycle or the input end or both, the mechanical capibility to make rapid queries to the file for the purpose of improving the quality of a retrieval (either directly or by improving the index) is essential to a system so large and diverse in coverage. The alternative is necessarily too many records or documents with the consequence that the analysts do not get what they need from the system and then begin to ignore it.

4.1.3 Types of Equipment

4.1.3.1 Index Transcription

The minimum requirement for this class of equipment is an electric typewriter, that is, one with the standard typewriter keyboard for entry of words and numerical tags. The typewriter must produce some form of recordable signal, such as a punched paper tape, magnetic tape, or a signal which is transmitted directly to a computer for storage there. The more elaborate devices give more service. For example, a magnetic-tape producing machine exists which simplifies correction of the tape at the local station. Direct connection to the computer gives quick service in the form of machine detection of errors and rapid feedback to the indexer. In some cases a combination of two typewriters might be desired in order to permit

the user to ask questions on one, receive an answer on the other, without breaking up the visual image of his text on the input machine. In other words, the indexer may choose to query a system for a given name and may not wish to disrupt the clean copy of his index with the listing of all information about that name.

4.1.3.2 Machine Readable Input

For machine readable copy, a conventional keyboard could be used but possibly a special, simplified keyboard would be better. The point to consider is that when machine readable copy is available, a great deal of transcription time can be saved by entering only word numbers or other word identifications rather than recopying the entire word. It would be desirable to permit the indexer to have one hand free to assist him in following the copy while using the other hand to operate the input keyboard. Thus, a simplified keyboard would be desirable.

4.1.3.3 Conversational Querying and Indexing

The first impression, when considering equipment for this task, is a cathode-ray tube console with multiple switches and possibly a light gun. This would present information to be displayed rapidly. Decisions could be made and then communicated rapidly. The excessive cost of such equipment precludes further consideration.

The same equipment proposed for indexing with a direct connection to a computer could do the job of conversational querying very well. That is, an electric typewriter keyboard with a direct connection and possibly some special keys to indicate special functions.

This system is slower than a cathode-ray tube, having a maximum of about 1,000 characters per minute against several thousand per second on a cathode-ray tube. However, the logical functions are the same and the typewriter has the advantage of providing copy much easier to read. That is, there is no flicker problem and the hard copy is preserved so that the analyst may retain several pages at his desk. Cost considerations are overwhelmingly in favor of the typewriter system, the cost ratio being on the order of ten to one, cathode-ray tube over typewriter.

4.1.3.4 Equipment Evaluation Problem

The overall system would benefit by having essentially the same device in use for all remote input functions whether query or indexing. Requirements in common are: typewriter keyboard, special function switches, optional ability to connect directly to a computer, optional possibility of linking two systems together for use by one person.

In addition, system designers may consider these devices as communications terminals, that is, as elements of the communication system. Buffering, then, is a consideration in selection of equipment. Use of the transcriber as a buffer will be considered in the next section, on communications.

4.2. Communications

4.2.1. The Communication Problem

Although this system is being designed to operate within a large building there is a communication problem in bringing data to the computer system and back again to the users.

In the presently operating systems, data is hand carried from room to room. In this system, much of the information flow will be electrical in order to meet possible requirements for on-line indexing and quick-response, conversational querying.

The elements required are I/O terminals, largely being modified electric typewriters; lines, here probably telephone or teletype; and multiplexers for switching and buffering. Buffering in such a system accomplishes two purposes. First, it holds information while awaiting processing. For example, if the computer is busy when a query arrives, it can be stored in a buffer until the computer is ready to take it. The second purpose is to store data temporarily for possible local processing before passing on to another communication point. The prime example of this is holding data at an indexer station until he is sure it is error free or to permit a second pass at the index record for adding additional data.

The problems faced in the design of this system are selection of equipment configurations to perform these basic functions and deciding where in the system each function will be performed.

4.2.2 Terminals

The I/O Terminals have been described under Input. Here, it should be pointed out that these devices can also act as buffers. There are three general classifications of terminals as buffers: no terminal buffering as in the case of a direct input device such as teletypewriter; paper tape buffering such as a Flexowriter gives to provide input sequencing flexibility and thus prevent communication

system overloading; magnetic tape buffering as in the message composer which gives communication flexibility and also permits rapid recall of records for error and correction.

There are other kinds of terminals also. In many systems there must be a device between the transcriber and the line. These generally perform some signal conversion function. Examples are the IBM 1051 or the AT and T Dataphone. Still another class of terminals is the security terminal which safeguards the signal. This device can affect the choice of the others.

4.2.3. Lines

The most likely choices here are telephone or Teletype lines. For security reasons, the latter seems the more appropriate for this project. No major implementation problems are expected in this area.

4.2.4 Multiplexers

These vary considerably in capability from simple switching to switching, buffering, and stored-program processing of data, including editing and priority determination. The simple switching multiplexers are virtually the same as terminals; the more elaborate ones are full-fledged computers.

Problems in choosing multiplexers mainly revolve around where to provide buffering capability, priority or overload procedures, and complexity of the communication problem presented (e.g., conversational querying and indexing is a much bigger communication problem because of the large number of duplex channels needed.)

4.2.5 Data Control

While not strictly speaking a communication problem, this is an appropriate place to introduce the concept of a Data Control Group. Its function would be control of files, especially new entries and changes to Authority and Fact Files. File maintenance is a complex and important process not only mechanically but also with respect to intelligence production. It may well be the decision of management that individual analysts may not make file modifications without proper approval and at first this requirement appears incompatible with the concept of high speed processing of data.

The general needs for a data control group would be the following: to have all prospective file changes routed to it before posting but not necessarily before actual entry into the computer and to have all files available for checking specific items or for general surveillance. In the sense of these requirements, the control group is simply another analytic group needing the same kind of access and communications with the computer as substantive analysts and indexers.

Hence, the concept of an analyst wishing to make direct postings to an information file and management wishing to exercise tight quality control over file entries are by no means incompatible.

4.3 Central computer complex

4.3.1 Computer

A high speed central processor of a type well within the current state of the art is definitely indicated. The computer must be capable of input-output over many channels in an overlap mode.

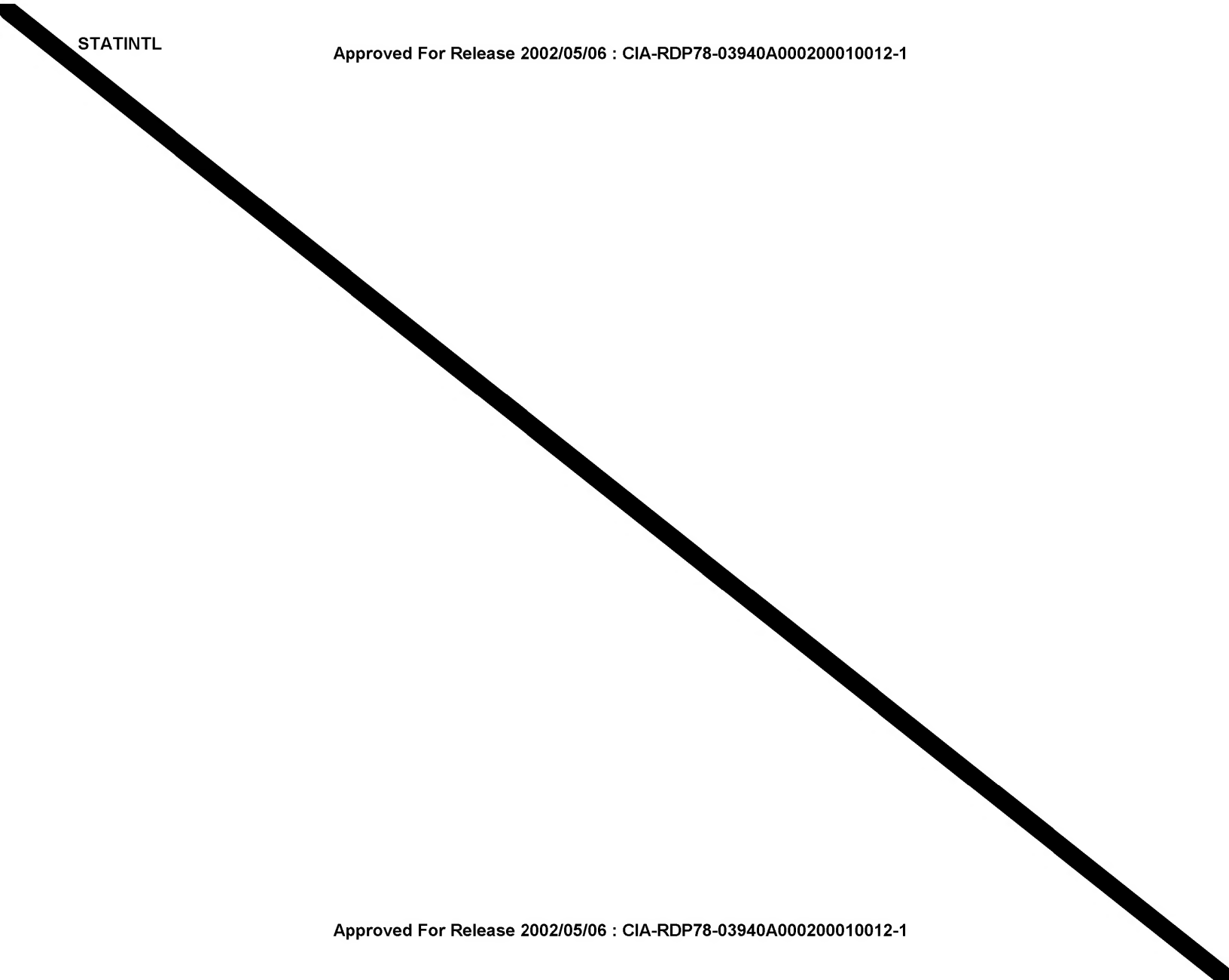
Processing speed should be high, but performance of the required searches in the required times should not present any major equipment problem.

4.3.2 Memory--

Memory capacity requirements implied by the system performance specifications are indicated in Figure 4-1. The three curves represent the fastest, slowest, and medium system implementation schedules. The capacity requirement presents a problem to the system engineering task. Random access devices such as disc storage devices are required. Several such devices are being considered which will satisfy the minimum system load requirements.

4.3.3 Programs

The possibility exists that the central computer complex will be time shared by the information retrieval system and other computer applications. Even without this possibility, the complexity of the information retrieval system itself necessitates the development of a control program. All of the functions of this control program cannot be given at this time. The file processing programs must be not only generalized but efficient. One of these is usually gained only at the expense of the other, but the proper balance must be attained for this system. Generalization is imperative in order to perform such requirements as cyclic queries and file interaction queries. In order to meet the timing specifications of the system, the programs must operate quite efficiently. Input-output programs must be written to process the traffic to and from the various on-line



Approved For Release 2002/05/06 : CIA-RDP78-03940A000200010012-1

stations. Process interrupt capability and priority handling capability are needed.

4.4 Document storage

4.4.1 Capacity

A similar graph to the one illustrating memory capacity requirements is shown in Figure 4-2 for document storage requirements. Again, the three curves represent the fastest, slowest, and medium system implementation schedules.

4.4.2 Search times

In order to satisfy the requirement for 15 minute document retrieval responses the document handling must be at least in part mechanized. Out of the 15 minutes, it is assumed that 10-12 minutes are available for document retrieval.

4.4.3 Document Form

Due to the sheer bulk of the document store for the system it is desirable to store microimages of the original hard copy documents. The actual retrieved item may be the microimage, a paper copy reproduced from the microimage, or a paper copy reproduced from the original document.

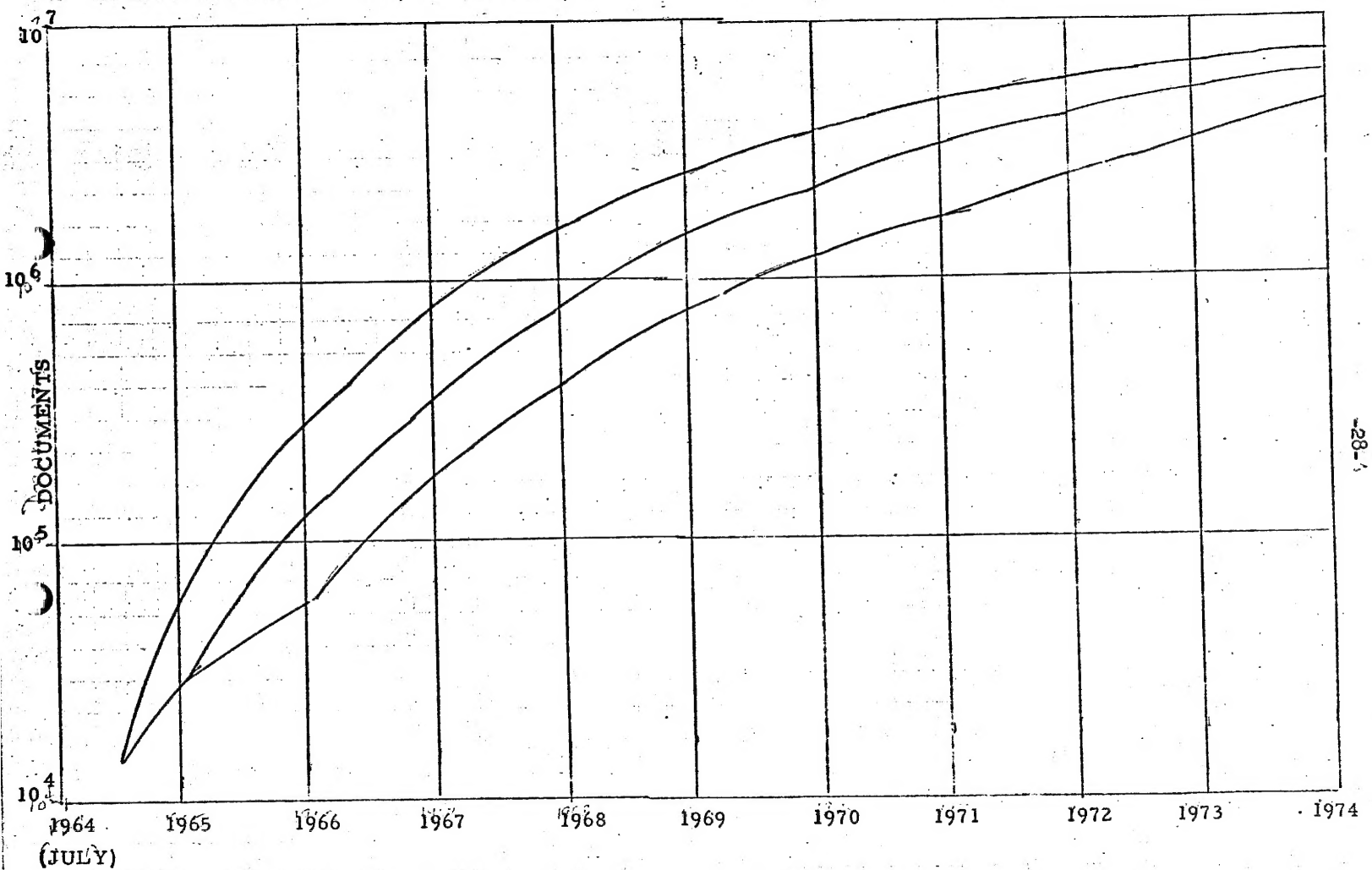


FIGURE 4-2

Document Storage Requirements